STATISTICS IN MEDICINE, VOL. 8, 201–216 (1989)

A METHOD FOR THE ESTIMATION OF CHRONIC DISEASE MORBIDITY AND TRENDS FROM MORTALITY DATA

ARDUINO VERDECCHIA AND RICCARDO CAPOCACCIA
Istituto Superiore di Sanita', viale Regina Elena 299, 00161 Rome, Italy

AND

VIVIANA EGIDI AND ANTONIO GOLINI
Dipartimento di Scienze Demografiche, University of Rome, v. Nomentana 41, 00161 Rome, Italy

SUMMARY

Measures of chronic degenerative disease diffusion, such as incidence and prevalence rates, are a basic need for epidemiologists and others working in many fields of human sciences. Equations relating death probabilities to incidence and survival probabilities for chronic degenerative diseases are derived from a cohort point of view. A maximum likelihood approach is adopted for the estimation of incidence as a function of time related covariates. When time series of mortality data are available, the model can be used to describe and analyse levels and dynamics of morbidity. A trial application to lung and breast cancer is given for the province of Varese, Italy, where incidence data are available from the Lombardy Cancer Register.

KEY WORDS Incidence Prevalence Chronic diseases Epidemiological methods Lung cancer Breast cancer

1. INTRODUCTION

Population health status at a certain time, like many other population structures, is the result of several phenomena acting together on the population throughout an extensive period. People are in time subjected to several risks that may change their health status (such as contracting disease, recovering, or dying from the same disease or from some other cause). All these risks are changing with time. For example, a decreasing mortality rate from a specific chronic disease could be attributed to different phenomena, such as a generalized reduction of the primary risk of the disease, or an improvement of diagnostic techniques and therapies which lead to a lower probability of dying or at least to a delayed death. Discrimination between such different possibilities can only be carried out through a comparative analysis of mortality, incidence, and prevalence data. Also, studies aimed at identifying the determinants of mortality and mortality differentials, at predicting mortality trends as well as the corresponding demand for health resources,[1] would profit greatly from a clear understanding of morbid processes and their dynamics.

While mortality data are usually available from routine statistics, incidence and prevalence rates can in general be obtained only from disease registers or epidemiological surveys for selected

0277–6715/89/020201–16$08.00

Received May 1987
Revised July 1988

pathologies. Such data are then usually limited in scale and rarely available at national or regional levels. Methods which allow estimation of incidence and prevalence rates from routine mortality data can be a great help when direct observation of incident cases is problematic or simply not undertaken. Deterministic models for the estimation of morbidity rates were developed at the International Institute for Applied Systems Analysis, Laxenburg, Austria. For chronic degenerative diseases which may be considered irreversible, Klementiev[2] proposed a model for estimating morbidity rates from mortality data suitable for application to steady morbidity processes in stationary populations. When applied to various diseases in Italy[3,4] this model was found to produce fairly good results. Kitsul[5] proposed a new model using a cohort approach, which no longer incorporated stationarity assumptions. However, in practical applications, when available mortality series are limited in time, stationarity of the morbid process is, in effect, still assumed in the proposed solution. The application of both models is limited mainly to highly lethal diseases.

Stochastic models of illness–death processes were developed theoretically and discussed extensively by Chiang.[6] Models of carcinogenesis have been developed and applied to the analysis of the age trend of mortality and incidence rates for selected cancers.[7–10] Using stochastic compartment models, Manton and Stallard[11–13] estimated annual onset and death rates for lung and stomach cancers, as well as the corresponding morbidity distribution for the population, from long series of cohort mortality data. For this purpose they formulated a bio-actuarial model which incorporated features of the human theory of carcinogenesis. However, no successful attempts are known for the modelling of diseases with complex age trends, such as cancer of the breast and uterus, or non-invasive diseases of the digestive system, diabetes, or cardiovascular disease for which it is more difficult to hypothesize a physiological model than for cancer.

An alternative approach consists of using information on the survival of diseased people to provide a functional description of the morbid process in an objective way but without making any assumption about the underlying physiological process. Survival data may be available from clinical follow-up of patients, epidemiological studies, or disease registries, and they can be found or collected for several diseases.

Following this type of approach, we propose a model which can be applied to non-stable populations, for estimating incidence and prevalence of chronic degenerative diseases, including those with low fatality rates or those which are uncommon in the population concerned.

2. MODEL FORMULATION

Assume for chronic degenerative diseases that the morbid process is irreversible. That is, recovery is not possible and an individual who becomes ill at a certain time remains ill until death.

Consider the compartment model of Figure 1, with two live states (healthy and sick with respect to the specific disease) and two death states (from the specific cause and from all other causes). Notice that only transitions relevant for formulating the model are shown in Figure 1.

From demographic sources we usually know death hazard rates from the specific cause, $\gamma(x)$, and from all causes together, $\alpha(x)$, as functions of age x. In addition, we may know from epidemiological sources the all-cause death hazard at age x, $\beta(x, y)$, for people who became ill at age y. Sometimes also the corresponding specific cause death hazard $\delta(x, y)$ may be known. When this is not true, the unknown intensity, $\delta(x, y)$, could be estimated by assuming independence of the death risks. In the following we assume that both transition intensities $\beta(x, y)$ and $\delta(x, y)$ are known.

Consider the ageing process acting on people in a defined cohort and let $\alpha(x)$, $\beta(x, y)$, $\delta(x, y)$ and $\gamma(x)$ indicate known transitions for this cohort. Our aim is to estimate the unknown transition

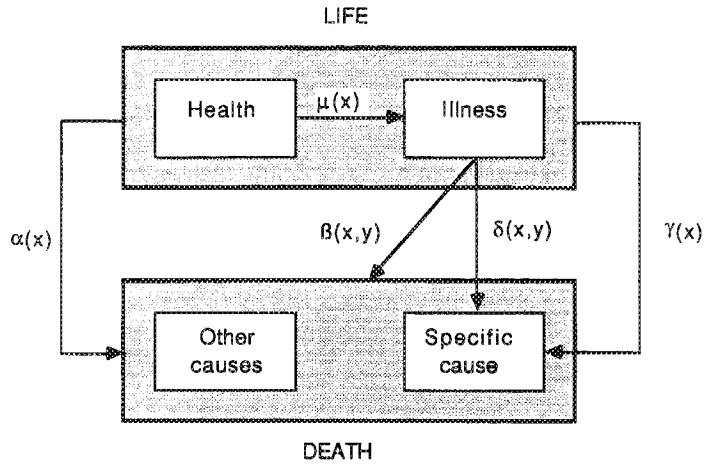

Figure 1. A compartmental representation of irreversible illness–death processes (see Section 2 for an explanation of the symbols)

intensity $\mu(x)$, that is, the disease hazard for healthy people of age x, as a function of the other known death hazards.

The survival probability $S(x)$ is defined for the cohort by:

$$S(x)=\exp\left[-\int_0^x \alpha(u)\mathrm{d}u\right]. \tag{1}$$

Define $\nu(x)$ as the probability of being in the sick state at age x, conditional on survival until age x. Then the death density function for the specific cause is given by:

$$S(x)\gamma(x)=\int_0^x S(\tau)[1-\nu(\tau)]\mu(\tau)\delta(x,\tau)\exp\left[-\int_\tau^x \beta(u,\tau)\mathrm{d}u\right]\mathrm{d}\tau. \tag{2}$$

Substituting $S(x)$ as defined by equation (1), we obtain:

$$\gamma(x)=\int_0^x [1-\nu(\tau)]\mu(\tau)\delta(x,\tau)\exp\left\{-\int_\tau^x [\beta(u,\tau)-\alpha(u)]\mathrm{d}u\right\}\mathrm{d}\tau. \tag{3}$$

Equation (3) expresses the observed cohort-specific death rate at age x as the integral over all ages up to x, of the probability of becoming ill at each age τ, times the probability of surviving the extra death risk for the diseased between ages τ and x, times the specific disease death density function at age x.

Similarly, the probability of being in the sick state for people of age x in the cohort can be expressed as:

$$S(x)\nu(x)=\int_0^x S(\tau)[1-\nu(\tau)]\mu(\tau)\exp\left[-\int_\tau^x \beta(u,\tau)\mathrm{d}u\right]\mathrm{d}\tau$$

which reduces to:

$$\nu(x)=\int_0^x [1-\nu(\tau)]\mu(\tau)\exp\left\{-\int_\tau^x [\beta(u,\tau)-\alpha(u)]\mathrm{d}u\right\}\mathrm{d}\tau \tag{4}$$

expressing the probability of being sick from the disease, for people aged x, as the integral over all younger ages of the probability of becoming sick at each age τ, times the probability of surviving the extra death risk between ages τ and x.

Equations (3) and (4) completely describe the morbidity and mortality of a cohort and can be used to estimate incidence and prevalence of the disease on the basis of complete cohort mortality data.

In practical applications, however, we generally deal with discrete statistics, grouped by age and calendar year, as they are recorded in vital registers. It is therefore convenient to express equations (3) and (4) in discrete terms suitable for use jointly with observed mortality and survival statistics. For this purpose we assume that diagnosis and death can occur only at the mid-point between two consecutive birthdays, except for those who become sick and die within the same year and to whom an average disease duration of six months is assigned. Under this assumption, we re-define in discrete terms some of the symbols above. We define v_i as the proportion of sick people at exact age i; μ_j as the probability of being diagnosed a new case for the specific disease between ages j and $j+1$ for a healthy individual at exact age j; d_{ij} as the probability of dying from the specific disease between ages i and $i+1$ for a sick person diagnosed between ages j and $j+1$ and who survived to exact age i, and γ_i as the probability of dying from the specific disease between ages i and $i+1$ for people who were alive at exact age i. Also define the discrete probability of surviving the extra death risk from the disease for sick people as:

$$s_{ij}=\exp\left\{-\int_{j+1/2}^{i}[\beta(u,j+1/2)-\alpha(u)]\,du\right\}.$$

Equations (3) and (4) respectively, can now, be rewritten in discrete terms as:

$$\gamma_i=(1-v_i)\mu_i d_{ii}+\sum_{j=0}^{i-1}(1-v_j)\mu_j d_{ij}s_{ij} \tag{5}$$

$$v_i=\sum_{j=0}^{i}(1-v_j)\mu_j s_{ij} \tag{6}$$

where the first term on the right side of equation (5) represents the contribution of people becoming sick and dying within the same year of age.

2.1. Maximum likelihood estimation

For each cohort, provided we knew the cause-specific death probabilities (γ_i), the corresponding probabilities for sick people with respect to the duration of the disease (d_{ij}), and the cause-specific relative survival accounting for the extra death risk to which sick people are exposed as compared to the general population (s_{ij}), and starting from the obvious condition $v_0=0$, equations (5) and (6) enable us to estimate the complete cohort history and its health status structure by age. However, direct numerical solutions require complete cohort mortality data, which are not easily available in practice, and may have problems of stability when random fluctuations of mortality data are quite high. Disease hazard is expected to vary continuously between contiguous age groups or contiguous calendar years or cohorts. We assume here that the incidence of the disease is a regular function of age, and possibly of other covariates, and look for an estimate of incidence function parameters using the model equations (5) and (6) to fit observed mortality data.

Suppose we have available mortality and population data for I age groups and a series of calendar years. Let μ_{ip} indicate the incidence probability at each age i in year p as given by a function, $M(\mathbf{z}, \boldsymbol{\theta})$, of a vector $\mathbf{z}$ of covariates, including at least age and period or cohort, and of a

set of unknown parameters θ. If survival probabilities are also available, the function $M(z, \theta)$ completely determines specific mortality in all cells of the given dataset.

We can use equations (5) and (6) to express the cause specific death probability γ_{ip} in each cell (i, p), belonging to the $(p-i)$th birth cohort, as a function of incidence at all younger ages, given by $M(z, \theta)$, for the same cohort in the past. Adding the subscript p to the equations (5) and (6) to indicate calendar year, we allow the various cohorts to have different disease histories. Death probabilities γ_{ip} can then be computed by the equations:

$$\gamma_{ip} = (1 - \nu_{ip})\mu_{ip} d_{iip} + \sum_{j=0}^{i-1} (1 - \nu_{jq})\mu_{jq} d_{ijq} s_{ijq} \tag{7}$$

and

$$\nu_{ip} = \sum_{j=0}^{i} (1 - \nu_{jq})\mu_{jq} s_{ijq} \tag{8}$$

where $q = p - i + j$ indicates calendar year of diagnosis. Substituting the values of ν into the first equation, we can express γ_{ip} as a function of incidence and survival alone and, ultimately, as a function of θ:

$$\gamma_{ip} = C_{ip}(\theta).$$

The parameters θ can then be estimated as those best reproducing the observed mortality data. If the probability of observing Y_{ip} deaths from n_{ip} people at risk can be assumed independently to follow a Poisson distribution with expectation $n_{ip}C_{ip}(\theta)$, then the log likelihood is given by:

$$L = \sum_{ip} (Y_{ip} \log[n_{ip}C_{ip}(\theta)] - n_{ip}C_{ip}(\theta)) + \text{terms independent of } \theta$$

and the maximum likelihood equations are given by:

$$\frac{\partial L}{\partial \theta_k} = \Sigma_{ip} \frac{\partial C_{ip}(\theta)}{\partial \theta_k} \left(\frac{Y_{ip}}{C_{ip}(\theta)} - n_{ip} \right) = 0, \qquad k = 1, \ldots, K.$$

Exact values for the derivatives in the above expression can be computed in each cell by differentiation of equations (7) and (8), and employed in the fitting procedure. Under general conditions, the maximum likelihood estimates of θ can also be obtained by an iteratively re-weighted least squares procedure,[14] that is, by maximizing, with respect to θ, the expression:

$$\sum_{ip} w_{ip}(Y_{ip} - n_{ip}C_{ip}(\theta))^2$$

where the weights w_{ip} are the inverse of the variance of Y_{ip},

$$w_{ip} = (n_{ip}C_{ip}(\theta))^{-1}$$

and are recomputed at each new step of the procedure.

Once the estimated values, $\hat{\theta}$, are obtained, asymptotic estimates of their standard errors can be calculated from second derivatives of log likelihood.

Goodness of fit can be evaluated by the likelihood ratio statistic (LRS):

$$\text{LRS} = -2 \sum_{ip} Y_{ip} \log \frac{n_{ip}C_{ip}(\hat{\theta})}{Y_{ip}}.$$

The likelihood ratio test can be used to test the significance of inclusion of one additional parameter to k already in the incidence model from the difference $G^2 = \text{LRS}_k - \text{LRS}_{k+1}$ which is asymptotically distributed as χ^2 with one degree of freedom.

According to the nature of the morbid process, we can choose the incidence function $M(\mathbf{z}, \theta)$ by different strategies to incorporate available data into the most appropriate model for specific descriptive or explanatory purposes. Age and period or birth cohort covariates should always be included in the function M when studying the dynamics of the disease.

3. APPLICATION

In this section we test the reliability of the above method by comparing model estimates with observed data for cancer in the Varese province of Italy, where incidence data are collected by the Lombardy Cancer Register (RTL).

Lung and breast cancer were selected for this analysis because of their different diffusion, age dependency, and fatality. In particular, lung cancer is characterized by high fatality and a steep rate of increase with age, with high incidence levels for men, and very low ones for women. Breast cancer, on the other hand, is representative of a high-diffusion, low-fatality process with an irregular age trend which makes this cancer particularly difficult to model.

3.1. Materials

Comparison of observed and estimated new cases of cancer was performed using the following data:

(i) number of deaths by age from lung cancer (men and women, ICD: 161), from breast cancer (women ICD: 174), and from all causes (men and women) in the Varese province, Italy, from official routine statistics for years 1970 to 1980;
(ii) size of the resident population of the Varese province by sex and age, estimated on 1 January of years 1970 to 1980;[15]
(iii) average number of new cancer cases by sex, age and site in the Varese province as recorded by RTL cancer register[16] in the years 1976 to 1977;
(iv) relative survival curves for lung and breast cancer by sex, from the Geneva Cancer Register (Switzerland) for 1970–1983.[17, 18]

As cancer survival data are not yet available in Italy, the Geneva Cancer Register was chosen as an alternative source, and corresponding survival probabilities were assumed for cancer occurring in the Varese province. The choice was made mainly on the basis of both geographical proximity and health care facilities not markedly dissimilar between Geneva and Varese, one of the most affluent provinces in Italy.

Relative survival is defined as the ratio of the survival probability observed in a group of sick people and that expected in an identically structured group of healthy people with respect to the specific disease. The relative survival accounts for the net death hazard from the specific disease.

The Geneva survival curves were not age-specific, and survival probability was necessarily assumed constant over age. Hence:

$$s_{ij} = \sigma_{i-j}$$

where σ_{i-j} is the relative survival at $i-j$ years from diagnosis as given by the survival curve. The crude probability of cancer death for sick people d_{ij} is not specifically known and has been computed from relative survival σ_{i-j} and the probability of cancer death for the general population, assuming that the death hazard from other causes for sick people is equal to the death hazard for the general population. The survival curves are shown in Figure 2.

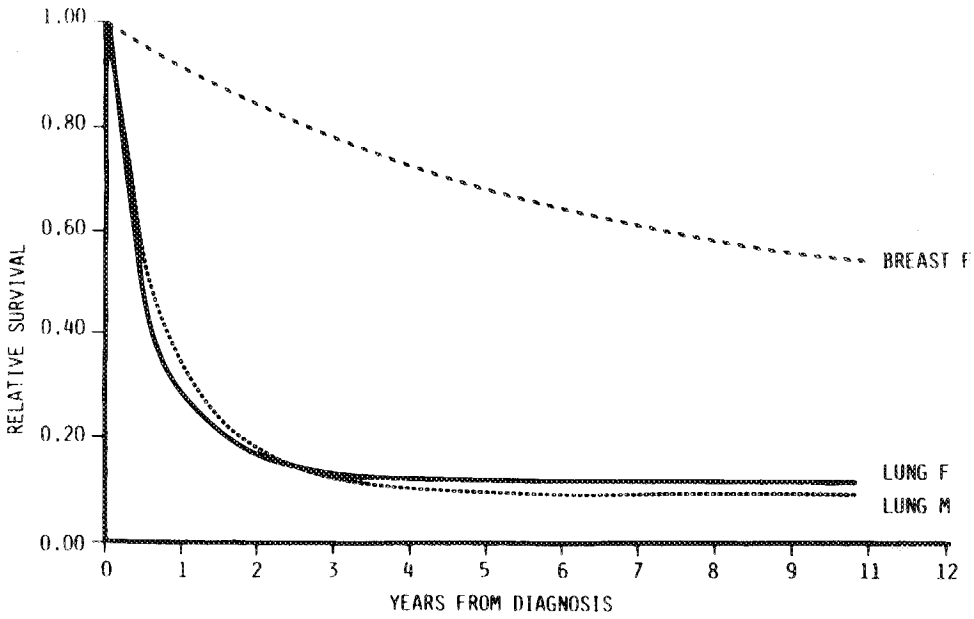

Figure 2. Relative survival curves for lung and breast cancer from the Geneva Cancer Register (years of diagnosis 1970–1977)

3.2. Maximum likelihood estimation of incidence and prevalence

The first step required for a solution by the method of maximum likelihood consists of identifying the type of function which is most appropriate for describing incidence of the specific disease.

In general, we can choose $M(\mathbf{z}, \theta)$ as a function of variables like age, period or cohort and eventually of other explanatory variables. Log-linear models are frequently used in connection with long-latency chronic diseases. In particular we choose a logistic function of age and birth cohort for use here with both lung and breast cancer. Higher powers of age and cohort have also been included to allow for non-linear relationships between variables and the logit of incidence. Formally:

$$\text{logit } M = a + \sum_{k=1}^{k1} b_k (\text{age})^k + \sum_{k=1}^{k2} c_k (\text{cohort})^k \tag{9}$$

where a, b_k and c_k are parameters to be estimated by the maximum likelihood method.

The orders $k1$ and $k2$ of the age and cohort polynomials in equation (9) must also be chosen. The procedure we used for this purpose is based on the likelihood ratio test applied iteratively to a series of increasing order nested models. The procedure terminates when the inclusion of further terms in equation (9) does not significantly improve the fit. The critical value for the G^2 statistic was set equal to 3·8, corresponding to a significance level of 0·05.

The steps in the verification of the incidence function for lung and breast cancer for the Varese province are reported in Table I.

The selection procedure described previously resulted in a third-order model, in both age and cohort for male lung cancer; a fourth-order model in age and linear in cohort for female lung cancer, and a third-order model in age with no cohort term for breast cancer.

Table I. Stepwise procedure for incidence model identification. The degrees of freedom for the likelihood ratio statistic (LRS) are given by the number of independent observations (935) minus the number of independent parameters

Site	Sex	Number of parameters	Order of polynomials		LRS	G^2
			age	cohort		
lung	men	2	1	—	1268·4	—
		3	2	—	574·9	693.5
		4	2	1	527·0	47·9
		5	2	2	507·9	19·1
		6	2	3	503·8	4·1
		7	3	3	497·6	6·2
lung	women	2	1	—	538·7	—
		3	2	—	516·4	22·3
		4	2	1	507·4	9·0
		5	3	1	503·6	3·8
		6	4	1	498·8	4·8
breast	women	2	1	—	906·5	—
		3	2	—	710·1	196·4
		4	3	—	669·3	40·8

The fit appeared quite good for all selected models. However, in evaluating the goodness of fit of the various models, it should be realized that data cells with zero observed deaths do not contribute to the LRS.[19] In our case, out of 935 cells, only 557 for male lung cancer, 353 for female lung cancer, and 576 for female breast cancer contributed to the reported LRS.

Age and cohort effects, as they result from the contribution of all related terms, can be described globally for the investigation of the structural and dynamic characteristics of the morbid process. In Figure 3, age trajectories of incidence for lung and breast cancer in Varese are depicted using the identified models. For breast cancer no cohort terms were found to be significant in the selected model and the age profile shown represents the actual estimated incidence. For lung cancer, which showed cohort trends for both sexes, the curves represent incidence values for the central 1943 cohort only. Incidence values for other cohorts can be obtained by shifting the 1943 curve upward or downward.

A double logarithmic scale is used in the plot to facilitate interpretation of incidence curves in terms of human carcinogenesis models. For lung cancer, the curves can be seen as a theoretical incidence age pattern adjusted for different risks associated with the cohorts, possibly owing to different smoking histories. The resulting relationship between log incidence and log age is not far from linear, being exactly linear for men aged 25 to 64 years. The corresponding slopes are 7·0 for men 25–74 (7·3 for men 25–64) and 5·5 for women 25–74. For breast cancer the trend is grossly non-linear, in accord with the well known relationship between risk of breast cancer and the three major events of reproductive life: menarche, first full-term pregnancy, and menopause.[20] It has been suggested[21, 22] that the deviation of the breast cancer incidence age curve from the 'log–log' linear pattern can be attributed to an exposure rate of breast tissue varying during lifespan according to age at these three events.

The dynamics of lung cancer incidence, incorporating all significant cohort terms, are presented in Figure 4. Here, relative risks of lung cancer incidence, along with 95 per cent confidence limits, are plotted against year of birth for cohorts 1885 to 1960. Relative risks were calculated with

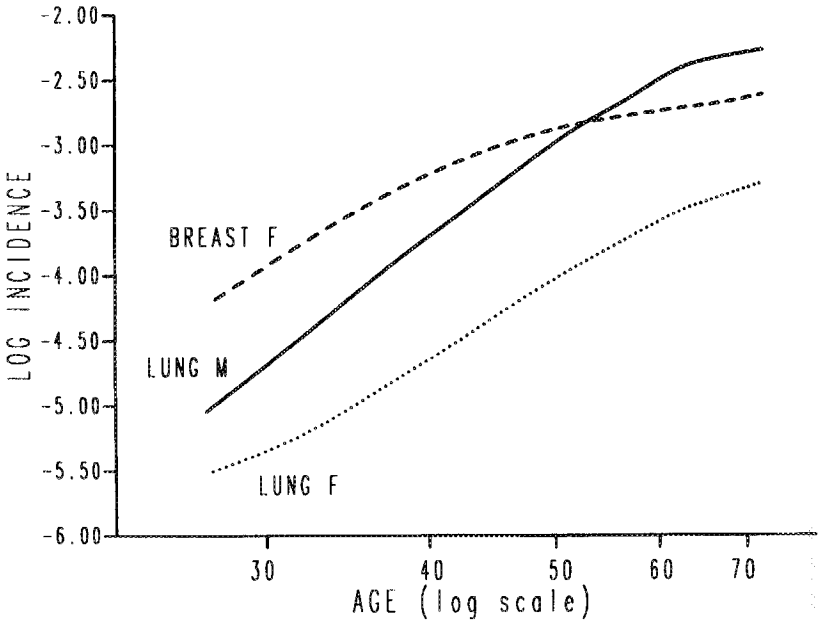

Figure 3. Age trajectories of incidence for lung and breast cancer

respect to the central cohort. The shape of the confidence lines results from the plotting of relative risks, that is, the ratio of two random variables. By definition, the relative risk for the reference cohort is one, with zero variance. The variance of the relative risk increases progressively away from the reference cohort as the covariance of the estimates decreases.

Lung cancer incidence shows an increasing trend from the earliest to the most recent cohorts, for both men and women. The rate of increase appears to drop for men, but not for women. Because of very small numbers of observed deaths in the cohorts from 1950 onwards the estimates of relative risk have large confidence intervals and no conclusions can be drawn. The estimated prevalences of the two types of cancer are reported in Table II.

For lung cancer, the ratio of prevalences for the sexes (men/women: 10) is approximately the same as those for incidence and mortality. On the other hand, breast cancer which is subjected to a weaker force of mortality, leads to a much higher proportion of prevalent cases than lung cancer. With a ratio of about 2:1 in the incidence of breast cancer and male lung cancer, the ratio between the corresponding proportions of prevalent cases is more than double at about 5:1.

Our results refer to a small area of Italy with a peculiar history, especially with regard to immigration, and therefore may not be representative of other areas or of the whole Italian population.

3.3. Comparison between estimated and observed cancer incidence

In Table III observed and estimated new cancer cases and incidence rates are reported for comparison. As a result of the absence of or very small numbers of cases in the youngest age groups and the unavailability of detailed reference data for ages over 75, only the intervening age groups are reported and compared.

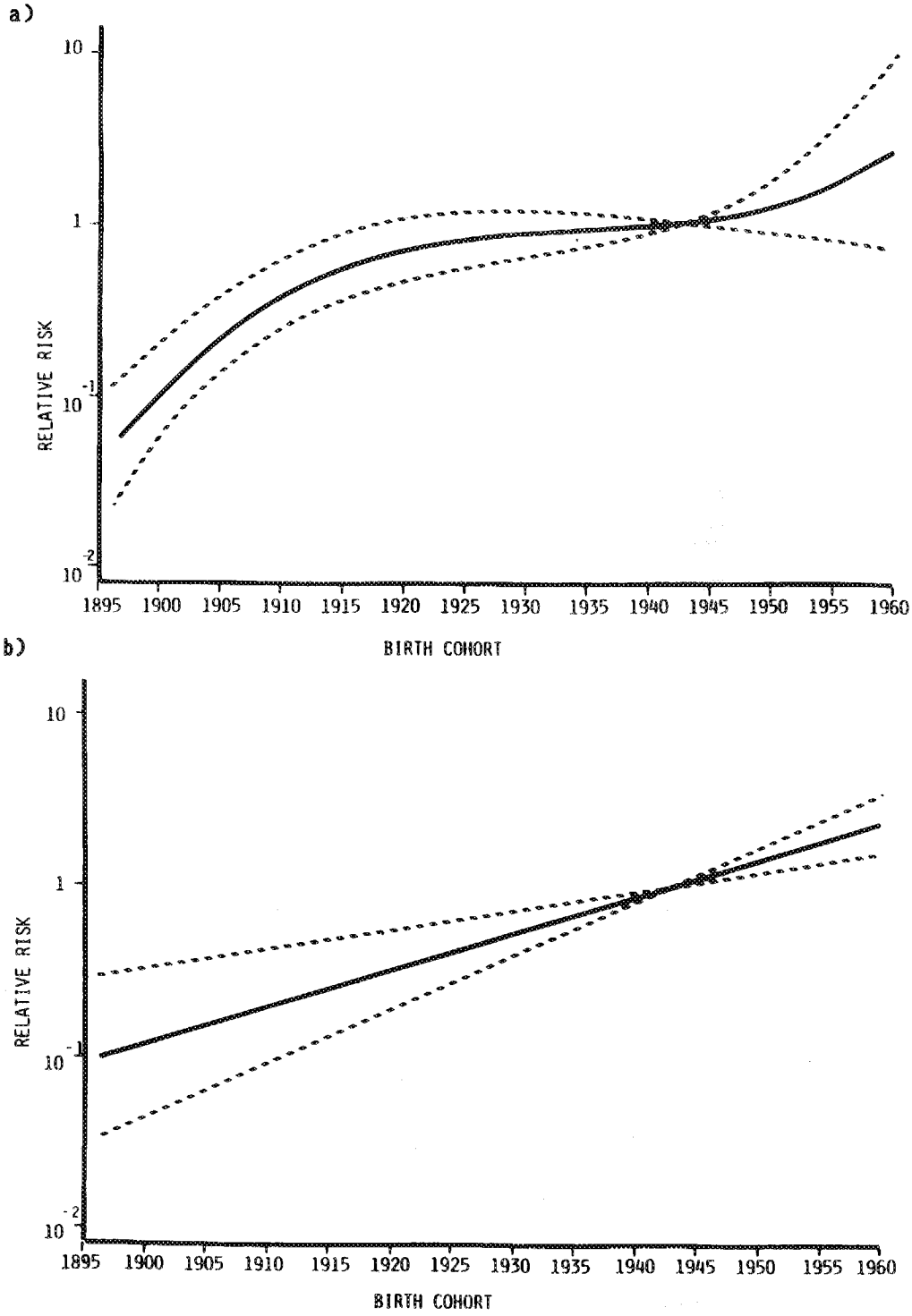

Figure 4. Lung cancer relative risk for cohorts born in the period 1885 to 1960 with respect to the 1943 cohort. Dashed lines indicate 95 per cent confidence intervals (a) men and (b) women

Table II. Population, number of prevalent cases and prevalence (× 10,000) of lung and breast cancer in Varese as estimated at 1 January 1977

	Males			Females				
		lung			lung		breast	
age (years)	population	cases	prevalence	population	cases	prevalence	cases	prevalence
25–29	30172	0	0·0	30421	0	0·0	5	1·6
30–34	27887	1	0·0	27698	0	0·0	16	5·8
35–39	30252	2	0·7	30355	0	0·0	47	15·5
40–44	27028	5	1·8	27066	1	0·4	93	34·4
45–49	25375	13	5·1	26131	1	0·4	166	63·5
50–54	22812	26	11·4	24947	2	0·8	251	100·6
55–59	15430	32	20·7	18026	3	1·7	246	136·5
60–64	16531	67	40·5	20401	6	2·9	401	196·6
65–69	15007	83	55·3	20520	9	4·4	481	234·4
70–74	10080	69	68·5	15816	9	5·7	455	287·7
25–74	220574	298	13·5	241381	31	1·3	2161	89·5

For male lung cancer, estimated incidence rates are in general agreement with the observed data. They slightly overestimate the observed cases in the older age groups.

For female lung cancer, even with very few deaths in each age group, the estimated incidence rates are generally in good agreement with the data except in a few age groups where an extraordinarily large or small number of cases were registered.

For breast cancer the model gave fairly good estimates, even though it failed to detect the two local maxima well evidenced by the observed incidence distribution in the age groups 45–49 and 55–59.

Unfortunately, no comparative data for prevalence are available from the RTL register nor from other sources. Our prevalence estimates, reported in Table II, appear plausible even if their reliability remains to be ascertained.

4. DISCUSSION

Our model can be applied to diseases other than those with short-term high-fatality or to those for which models of onset and evolution can be specified. The principal assumption in the formulation of the model was the irreversibility of the morbid process. For many chronic degenerative diseases, such as ischaemic heart disease, liver cirrhosis, or diabetes, this is plausible. For malignant neoplasms, it reflects the idea of cancer load, since a patient with a history of cancer is still expected to make greater use of health resources than healthy people never afflicted by the disease. A concept of cancer prevalence which would be closer to actual disease manifestation than the one adopted here is very difficult to define precisely. However, the introduction of the concept of recovery into the model remains an important task for future work, both to increase the field of application and to take into account possible progress in clinical management of chronic degenerative diseases.

The aim of this paper is to show how the relationships between incidence, survival, and mortality can be used either to check the consistency of observed data or to estimate incidence and

Table III. New cases and incidence (probability × 10,000) of cancer in Varese, in the two-year period 1976–77, as observed and estimated by the proposed model

Age	Observed cases	Observed incidence	Expected cases	Expected incidence	χ^2
(a) *male lung cancer*					
30–34	2	0·46	1·9	0·34	0·01
35–39	3	0·50	6·4	1·10	1·81
40–44	21	3·88	15·4	2·84	2·04
45–49	39	7·66	33·5	6·58	0·90
50–54	66	14·40	61·4	13·40	0·34
55–59	77	24·76	70·6	22·70	0·58
60–64	117	34·91	125·6	37·48	0·59
65–69	133	43·38	144·3	47·07	0·88
70–74	102	48·94	111·4	53·45	0·79
(b) *female lung cancer*					
30–34	0	0·00	0·4	0·06	0·40
35–39	2	0·33	0·8	0·14	1·80
40–44	2	0·37	1·7	0·31	0·05
45–49	3	0·57	2·3	0·64	0·21
50–54	4	0·80	5·9	1·18	0·61
55–59	10	2·77	6·9	1·93	1·39
60–64	10	2·44	12·3	3·00	0·43
65–69	18	4·35	16·6	4·01	0·12
70–74	5	1·56	16·3	5·07	7·83
(c) *female breast cancer*					
25–29	4	0·66	3·6	0·59	0·04
30–34	6	1·08	9·7	1·76	1·41
35–39	24	3·95	24·7	4·07	0·02
40–44	43	7·94	39·9	7·36	0·24
45–49	82	16·05	57·8	11·05	10·13
50–54	68	13·60	71·7	14·33	0·19
55–59	77	21·29	60·4	16·70	4·56
60–64	81	19·74	77·4	18·87	0·17
65–69	91	21·98	86·4	20·89	0·24
70–74	71	22·10	77·8	24·22	0·59

prevalence in areas for which they have not been recorded. The model can be used for different purposes, according to the availability of reliable data, except that mortality data are assumed to be available. A number of distinct cases may be considered.

When incidence and survival are known, indirect estimates are unnecessary. Often, however, incidence, survival and mortality data are derived from different sources, not necessarily compatible with each other with regard to methods of collection, criteria of exclusion, and completeness of observation. In this case, equations (7) and (8) can be used to evaluate the degree of consistency of available data and to supply prevalence estimates.

Sometimes incidence and survival data are obtained from a register in a selected population (generally defined according to area of residence). If the population covered by the register is representative of the general population for available health care facilities (both diagnostic and therapeutic), then the observed survival data may be reasonably extended to the latter. Such extension is not generally valid for incidence, which is strongly affected by the geographical

variability of risk factors, related to both environment and life-style. The observed survival data can be used jointly with mortality rates to yield incidence and prevalence estimates in the population as a whole.

When incidence data only are available from the register, a survival curve for an external population comparable to that in the study at least for the standard of health facilities might be selected; the curve should be adjusted to account for the principal demographic variables. As a preliminary step, the validity of the selected survival curve must be ascertained by attempting to use it to reproduce the incidence observed from the register. If this is successful, the survival curve may be used, bearing in mind the cautions above, to estimate incidence and prevalence in the general population.

If no morbidity data exist for the disease under consideration, application of the method, with a plausible survival curve from another population, would give hypothetical estimates of morbidity levels. Their reliability, could not be verified because of lack of independent control data to support the assumptions about survival.

The reliability of the estimates obtained by applying the proposed model is strictly dependent on the quality of the data. For example, use of an excessively optimistic survival curve would lead to overestimation of both incidence and prevalence rates. Or again, incompleteness of the registration of incidence would lead to a corresponding underestimation of incidence in the general population, and so on.

It is well known that all morbidity and mortality measures in open populations present considerable problems; extensive discussion of this issue, which is widely treated in the literature (see, for example Reference 23) lies outside the scope of this paper. Ascertaining the quality and completeness of data with regard to possible sources of error and distortion should be a preliminary step in using the method proposed, even more so than in conventional descriptive analyses. Special attention should be devoted to the choice of survival data when not directly available for the population concerned. Reported survival rates can vary between populations just as between individuals, for different reasons which we can classify under three headings. First, we have the types of treatments and specific prognostic factors associated with objective differences in the duration of disease. Their effect is difficult to quantify, and depends greatly on the specific disease considered. For cancer, there is some indication that these differences are small, at least for certain sites such as lung, stomach and uterus.[24,25] Secondly, we may consider demographic factors, such as age, sex, and race, which can generally be explicitly accounted for. Finally, we consider apparent variations arising from different stages of disease at which incident cases are diagnosed, as they provide the basis for subsequent observation of survival. For cancer, apparently improved survival rates could arise from early diagnosis (lead-time bias). In cardiovascular disease, observed survival rates can depend considerably on the completeness of the registration process for cases of sudden death, not receiving hospital care. If the definition of incident case is changed to one holding in the population from which survival data are taken, the effect of these variability factors can no longer be considered as causing distortion, but rather a sort of standardization by differences in diagnostic standards. For example, assume that cancer cases are diagnosed in population A during the first (local) stage of spreading, while in population B they are observed only on entering into the second (regional) stage. Also assume that the survival differences between the two populations arise from the different starting points of observation (average time of passage from stage I to stage II). Then, using the survival data from population A on population B, we can obtain incidence estimates which correspond to those which would be observed if here, too, cases were observed at stage I.

Variability in survival rates is generally lower than that for incidence rates of many chronic diseases. In European countries, published data[26] report, for example, that the observed incidence

of cancer of the colon in men for 1975 varies from 5·5 per 10,000 (Romania) to 23·8 per 10,000 (North Scotland, U.K.); the incidence of lung cancer varies from 23·5 (Navarra, Spain) to 96·8 (West Scotland, U.K.); the incidence of cancer of the breast varies from 18·9 (Cieszin, Poland) to 76·1 (Geneva, Switzerland); the incidence of cervical cancer varies from 3·9 (Navarra, Spain) to 30·1 (G.D.R.). For coronary heart disease the 5-year incidence rate observed in the Seven Countries Study[2,7] varies from 793 per 10,000 in Finland to 164 per 10,000 in Greece. In all these examples the ratio of maximum and minimum values is at least equal to 4. The same variability attached to death hazard rates would be reflected in survival probabilities varying, for example, from 0·65 to 0·90, or from 0·06 to 0·50. Variations of this order are not realistic and cannot easily be inferred for countries with a comparable level of development. Moreover, comparative studies[24, 25, 28] have shown that, for many cancer sites, variability in survival rates between populations is quite low and surely lower than the variability of incidence.

As has been asserted on a number of occasions, the application presented in this paper is aimed at primarily illustrating and testing the method. No particular epidemiological significance is claimed for the results obtained. However, a number of assumptions made here, should be discussed briefly. Adopting relative survival curves from the Geneva cancer register in the application to lung and breast cancer, led us to assume survival probabilities independent of both age and period. For lung cancer it is known that survival for younger people is higher than for the elderly, both for males and females.[29, 30] This could be used to explain the model's tendency to overestimate, albeit not to any considerable extent, incidence in the oldest age groups. For breast cancer, the age dependency of survival probability is more complex.[29–33] In the United States, for example, 5-year survival rate for breast cancer for white females reaches a maximum (0·75) in the age group 45–54, declining to 0·71 for ages 65–74, with some evidence of slight decrease for very young age groups. However, this age dependency of relative survival rates for breast cancer is less marked than for lung cancer, at least if the analysis is confined to below age 75. As far as period is concerned, there are indications of an increase in cancer survival in recent years, particularly for breast cancer. This issue is still controversial,[34] and it is, therefore, difficult to make reliable assumptions about it.

The probability of death from causes other than the specific disease was assumed the same in sick people and in the general population. It is worth remembering that this assumption is not demanded by the model, but results from the lack of conclusive epidemiological data on multiple disease endpoints. It is necessary to remember that if a positive association existed (so that the diseased would be at a greater risk from other causes) this would lead to overestimation of the specific death hazard rate, and hence to underestimation of incidence. Conversely, a negative association would lead to a distortion in the opposite sense. For cancer, both these effects should be small. It is likely that the debilitating effect of cancer could result in an increased risk for other diseases (such as cardiovascular or respiratory disease). However, as internationally accepted rules for coding causes of death state that cancer should be taken as the initial cause of death even when it is mentioned as the second or the third cause on the death certificate, most of these cases will be indeed classified as cancer deaths. A positive association is likely to exist, even though probably weak in practice, provided that official death records are taken as the source of mortality data. A negative association can be the result of increased use of health resources by diseased people. The resulting bias is, however, small. For cancer patients by far the most relevant risk of death is cancer itself. Even with complete elimination of all other causes of death their survival rates would not be substantially increased.

Finally, in the comparison of estimated and observed cancer cases, possible under-reporting of new cases must be taken into account. In the Varese register this effect is unlikely to be important. The completeness of registration can be evaluated by means of the percentage of cases detected

through the death certificate ('death registration only' cases); the lower this value, the better the quality of register incidence data. In our case this percentage, during the period considered and for ages 0–74, was about 2 per cent for lung cancer in men and breast cancer, and about 4 per cent for lung cancer in women.

REFERENCES

1. Golini, A. and Egidi, V. 'Effects of morbidity changes on mortality and population size and structure', IUSSP Seminar on Methodology and Data Collection in Mortality Studies, Dakar, Senegal, 7–10 July, 1981.
2. Klementiev, A. A. 'On estimation of morbidity', International Institute for Applied System Analysis, Laxenburg, Austria, RM-77-41, 1977.
3. Verdecchia, A., Capocaccia, R. and Mariotti, S. 'Estimation of cardiovascular diseases morbidity using mortality data', *in* Eimeren, W., Engelbrecht, R. and Flage, Ch. D. (eds.) *Proceedings of Third International Conference on System Science in Health Care*, Springer Verlag, 1984.
4. Verdecchia, A., Capocaccia, R. and Mazzoni, C. 'Estimation of cancer morbidity using mortality data', *Tumori*, **71**, 431–439 (1985).
5. Kitsul, P. 'A dynamic approach to the estimation of morbidity', International Institute for Applied Systems Analysis, Laxenburg, Austria, WP-80-71, 1980.
6. Chiang, C. L. *Introduction to Stochastic Processes in Biostatistics*, Wiley, New York, 1968.
7. Armitage, P. and Doll, R. 'The age distribution of cancer and a multi-stage theory of carcinogenesis', *British Journal of Cancer*, **VIII**, 1–12 (1954).
8. Cook, P. J., Doll, R. and Fellingham, S. A. 'A mathematical model for the age distribution of cancer in man', *International Journal of Cancer*, **4**, 93–112 (1969).
9. Doll, R. and Peto, R. 'Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers', *Journal of Epidemiology and Community Health*, **32**, 303–313 (1978).
10. Day, N. E. and Brown C. C. 'Multistage models and primary prevention of cancer', *Journal of the National Cancer Institute*, **64**, 977–989 (1980).
11. Manton, K. and Stallard, E. 'A cohort analysis of U.S. Stomach Cancer Mortality 1950–1977', *International Journal of Epidemiology*, **11**, 49–61 (1982).
12. Manton, K. and Stallard, E. 'The use of mortality time series data to produce hypothetical morbidity distributions and project mortality trends', *Demography*, **19**, 223–240 (1982).
13. Manton, K. G. and Stallard, E. *Recent Trends in Mortality Analysis*, Academic Press, 1984.
14. Frome, E. L. 'The analysis of rates using Poisson regression models', *Biometrics*, **39**, 665–674 (1983).
15. Capocaccia, R. and Scipione, R. 'Stima della popolazione italiana per eta', sesso e provincia di residenza negli anni 1971–1979', *Epidemiologia e Prevenzione*, **21/22**, 57–71 (1985).
16. Berrino, F., Crosignani, P., Riboli, E. and Vigano', C. 'Epidemiologia dei tumori maligni: incidenza e mortalita' in provincia di Varese 1976–77', *Notizie Sanita*, **31**, (1981).
17. Registre Genevois des Tumeurs 'Survie des cas incidents de la periode 1970–77', Preliminary report, 1982.
18. Registre Genevois des Tumeurs 'Cancer a Geneve: incidence, survie, mortalite 1970/1983', 1984.
19. Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Mass., 1975.
20. Miller, A. B. and Bullbrock, R. D. 'The epidemiology and the etiology of breast cancer', *New England Journal of Medicine*, **303**, 1246 (1980).
21. Pike, M. C., Krailo, M. D., Henderson, B. E., Casagrande, J. T. and Hoel, D. G. ' "Hormonal" risk factors, "breast tissue age", and the age-incidence of breast cancer', *Nature*, **303**, 767–770 (1983).
22. Pike, M. C. 'Age-related factors in cancer of the breast, ovary, and endometrium', *Journal of Chronic Diseases*, **40**, Suppl. 2, 59S–69S (1987).
23. Doll, R. and Peto, R. 'Avoidable risk of cancer in the U.S.', *Journal of the National Cancer Institute*, **66**, 1192–1308 (1981).
24. Logan, W. P. D. 'Cancer survival statistics international data', *World Health Statistics Quarterly*, **31**, 62–73 (1978).
25. Hakulinen, T. 'A comparison of nationwide cancer survival statistics in Finland and Norway', *World Health Statistics Quarterly*, **36**, 35–46 (1983).

26. World Health Organization, and International Agency for Research on Cancer. 'Cancer incidence in five continents', Vol. IV, 1982.
27. Keys, A. *Seven Countries*, Harvard University Press, Cambridge Massachusetts, 1980.
28. Horner, R. D., and Chirikos, T. N. 'Survivorship differences in geographical comparison of cancer mortality: an urban-rural analysis', *International Journal of Epidemiology*, **16**, 184–189 (1987).
29. Axtell, L. M., Asire, A. J. and Myers, M. H. 'Cancer patient survival', Report No. 5, NIH Publication No. 81–992, 1976.
30. The Cancer Registry of Norway. *Survival of Cancer Patients*, Oslo, 1980.
31. Wilkinson, G. S., Edgerton, F., Wallace, H. J., Reese, P., Patterson, J. and Priore, R. 'Delay, stage of disease, and survival from breast cancer', *Journal of Chronic Diseases*, **32**, 365–373 (1979).
32. Dayal, H. H., Power, R. N. and Chiu, C. 'Race and socio-economic status in survival from breast cancer', *Journal of Chronic Diseases*, **35**, 675–683 (1982).
33. Hankey, B. F. and Steinhorn, S. C. 'Long term patient survival of the more frequently occurring cancers', *Cancer*, **50**, 1904–1912 (1982).
34. United States General Accounting Office (GAO). 'Cancer patient survival', report to the Chairman, Subcommittee on Intergovernmental Relations and Human Resources, Committee on Government Operations, House of Representatives, Washington, U.S.A., March 1987.